Identifying Biomarkers of Alzheimer's Disease via Integrated Principal Components Analysis

Jaihee Choi, Rice University
Ying-Wooi Wan, Baylor College of Medicine
Rami Al-Ouran, Baylor College of Medicine
Zhandong Liu, Baylor College of Medicine
Genevera Allen, Rice University

WOMEN IN DATA SCIENCE CONFERENCE

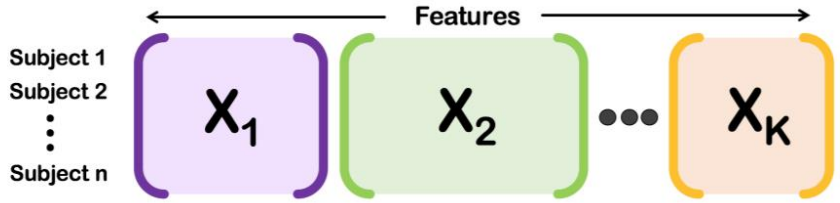
OCTOBER 23, 2020

Goal of this Project

- **Integrate multiple data sets** to identify meaningful clusters
- See patterns in clinical data from clusters formed
- Identify genes that are related to clusters of subjects with Alzheimer's disease

Summary of Methods

- Data: Religious Orders Study and Memory Aging Projects
  - miRNA (507 x 309)
  - RNA-seq (507 x 900)
  - methylation data (507 x 1250)
- Methods:
  - Integrated Principal Component Analysis [Tang, Allen (2018)]

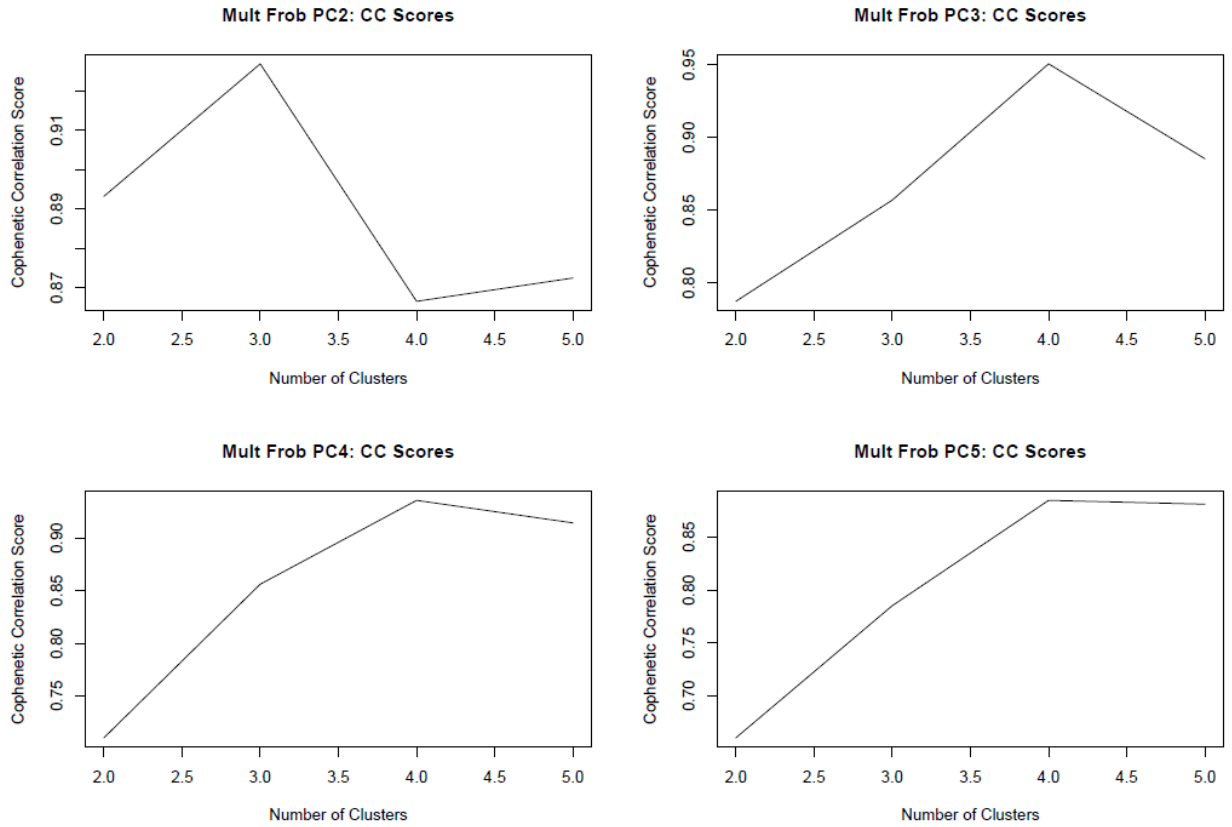


Integrated Data Setting for iPCA. From *Integrated Principal Components Analysis* by Tang, Allen (2018). Retrieved from https://arxiv.org/pdf/1810.00832.pdf.

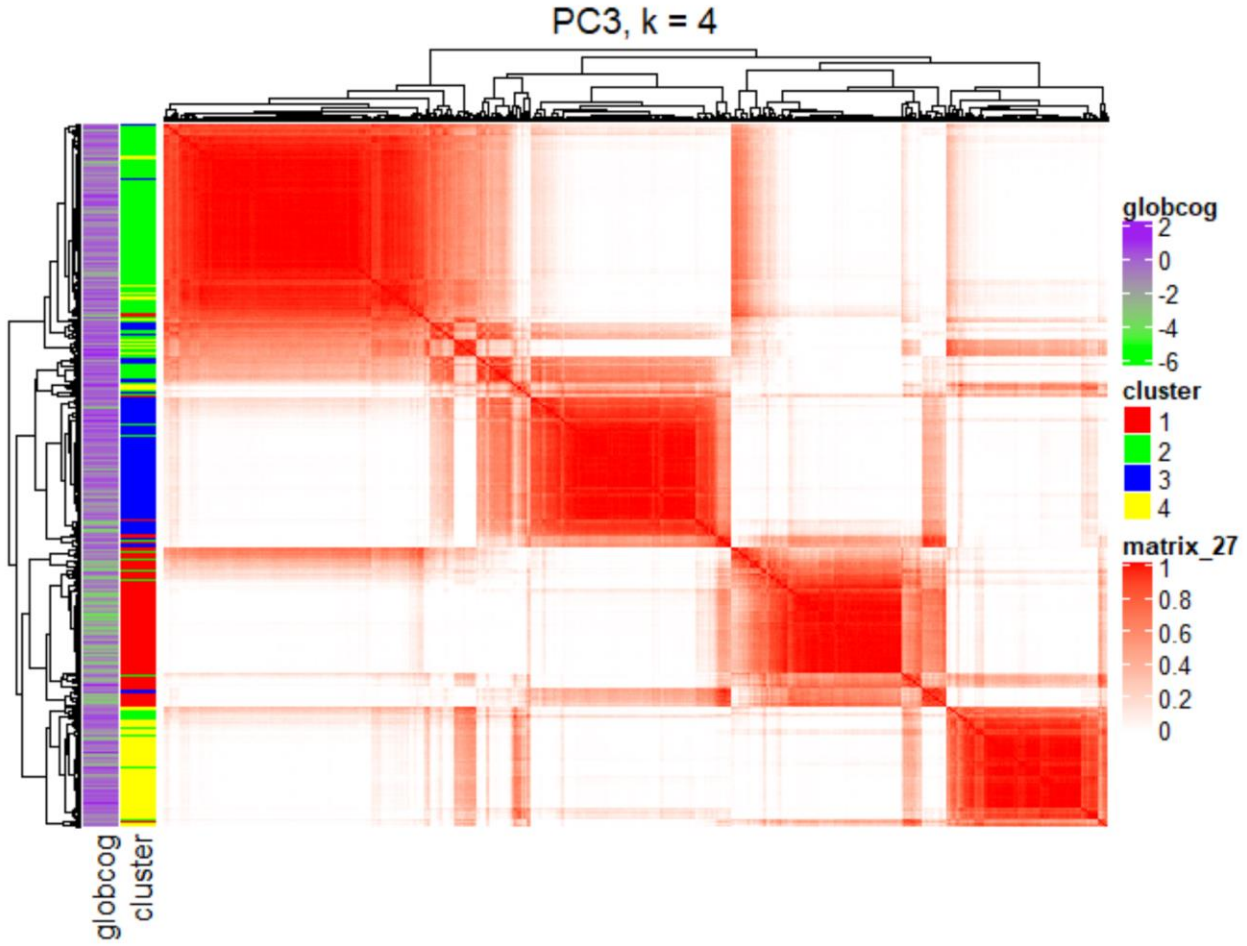
Identifying Clusters

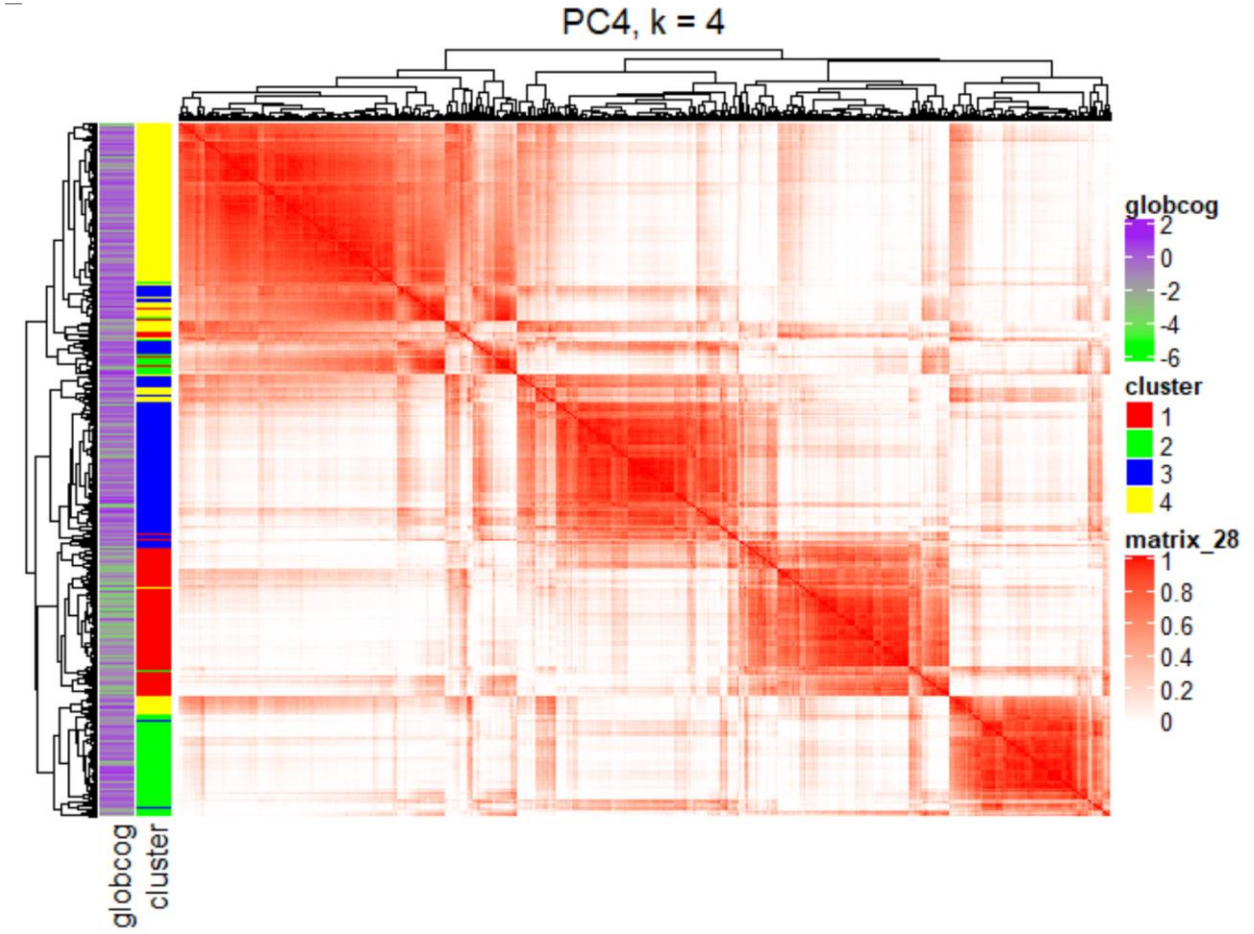
- **Goal:** Find robust clusters using clustering algorithms across different PC values
  - need to find number of PCs to use and how many clusters
- Use **Consensus Clustering** to see how stable clusters are
  - Calculate Cophenetic Correlation Coefficient [Brunet et. al (2004)]

Cophentic Correlation Score for Different Number of PCs

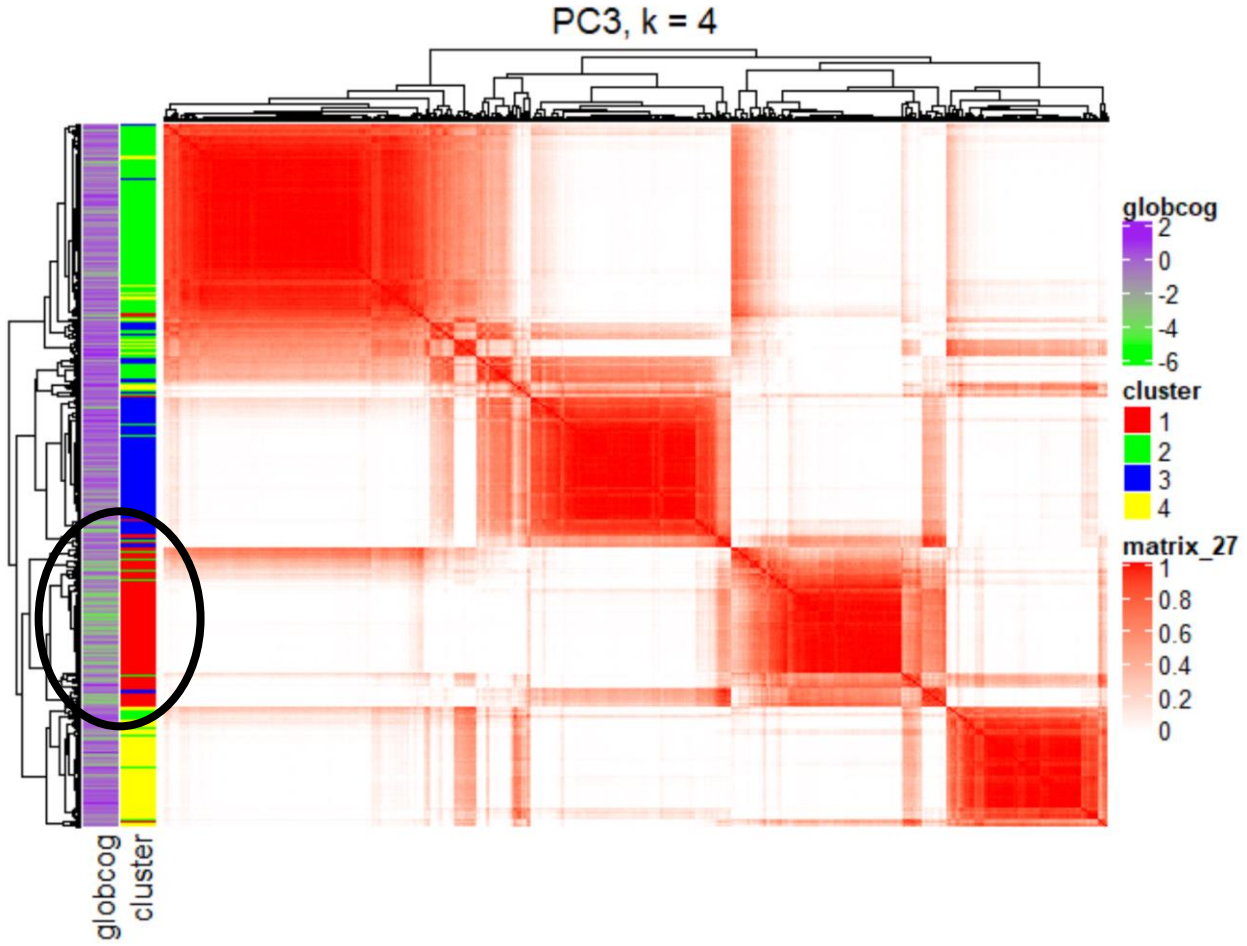


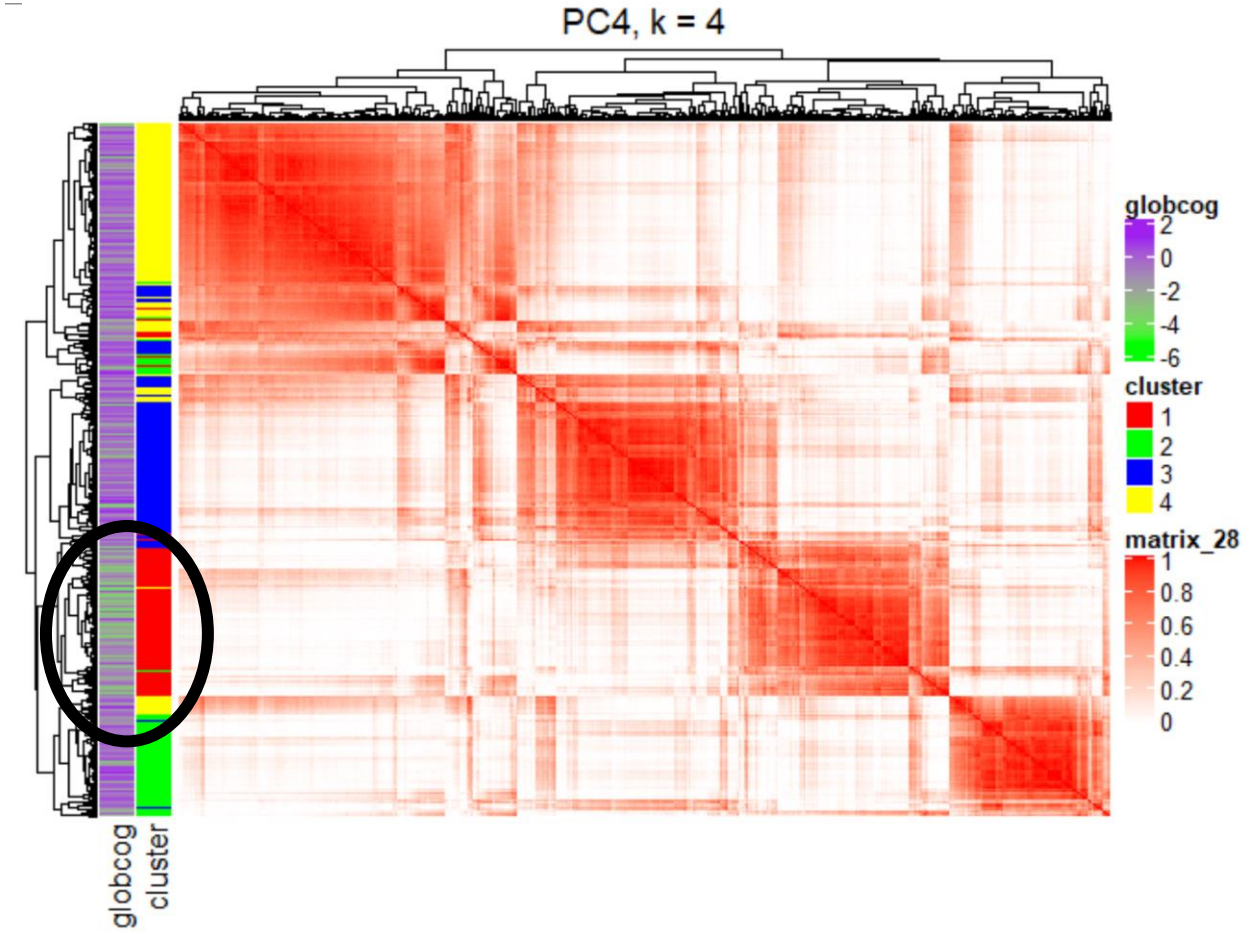
Heatmap of Consensus Cluster Matrix



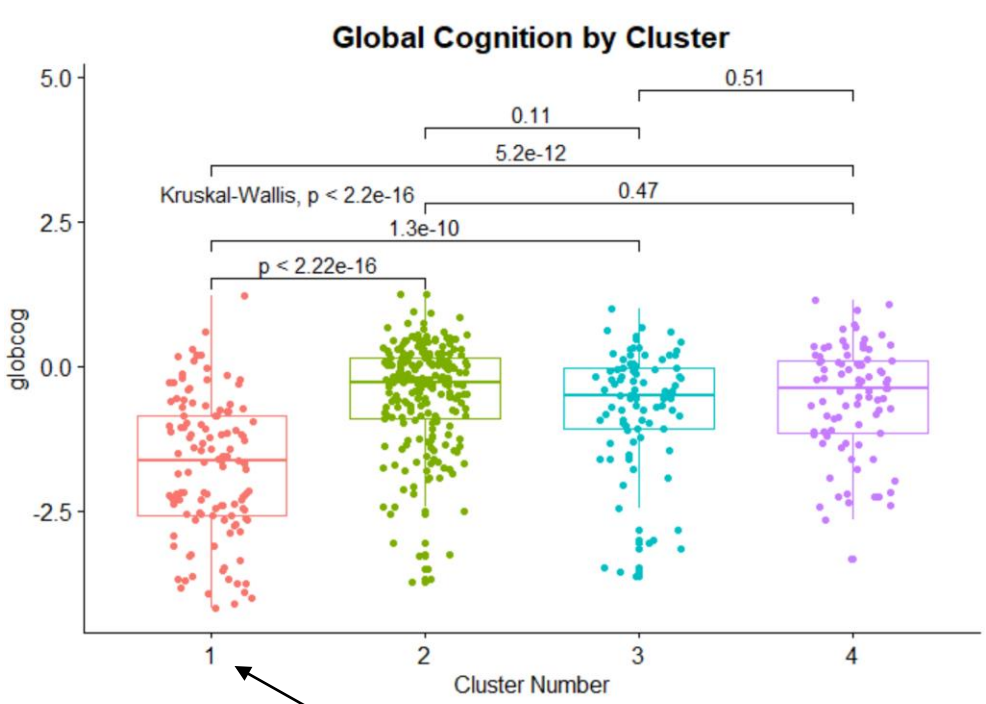


Heatmap of Consensus Cluster Matrix



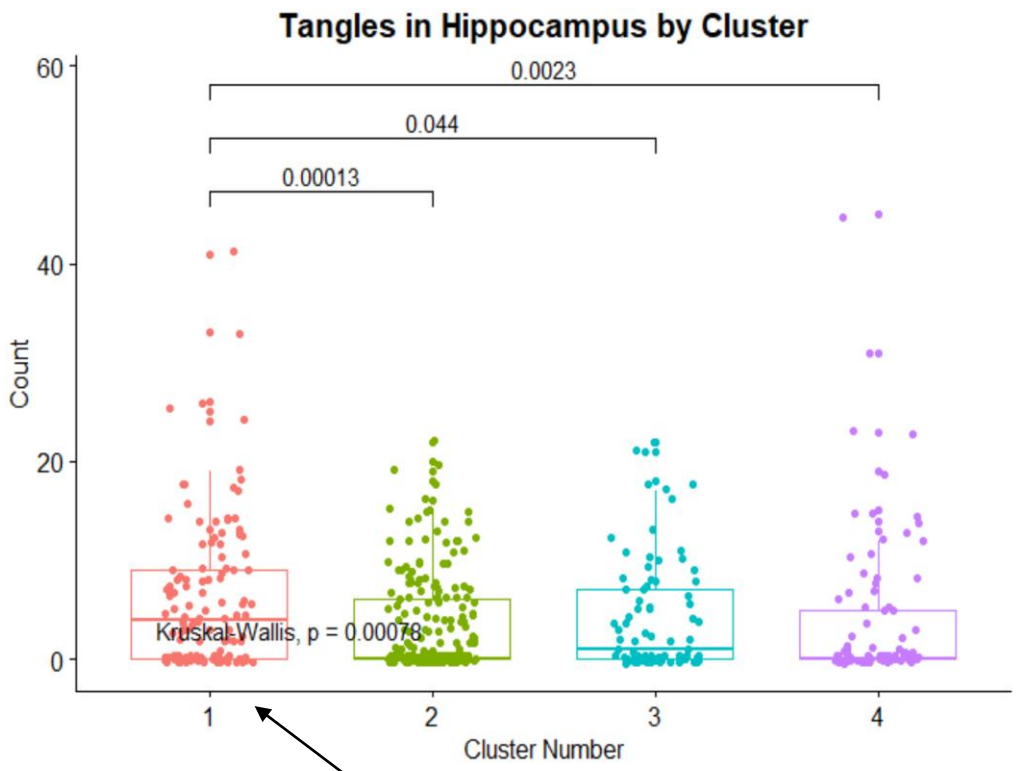


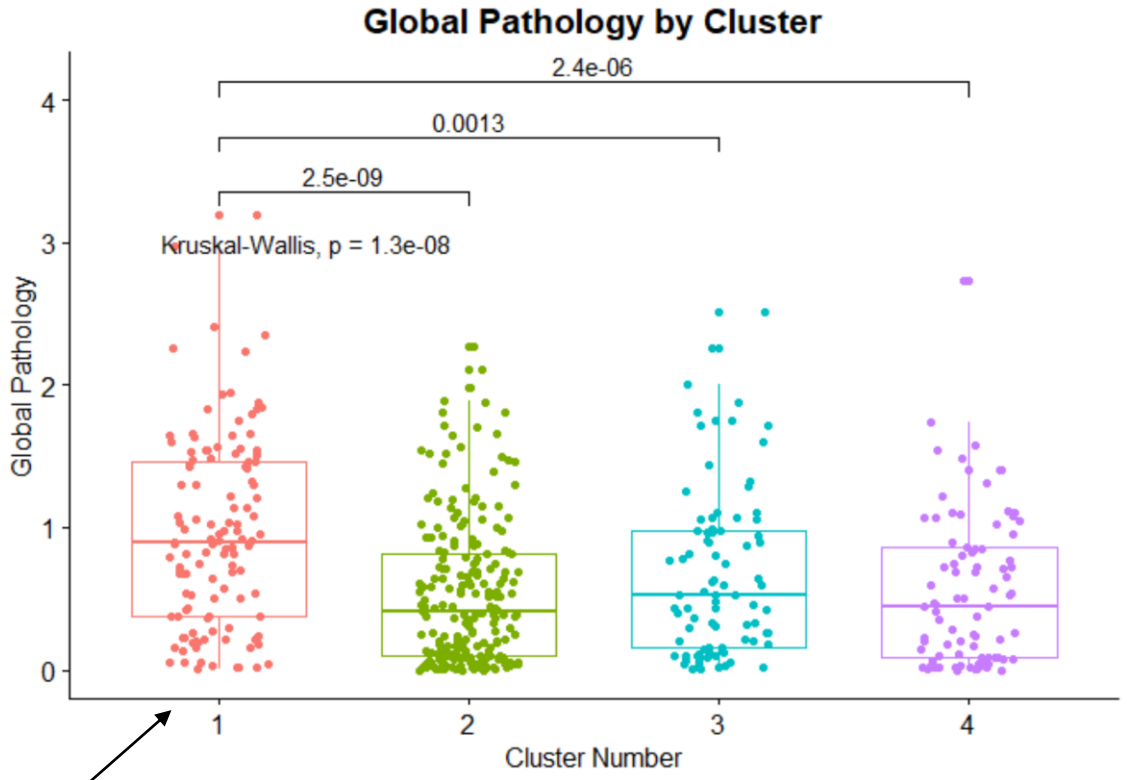
Significant Clinical Variables —PC: 1-3, k = 4



Cluster 1 has significantly lower values for both variables

Other Clinical Variables—PC: 1-3, k = 4





Cluster 1 has significantly higher values for both variables

Linking to Genes

1. Apply correlation test to PCs to view relationships
2. Filter out significant correlation between genes and PCs
3. Rank correlation of genes to PCs and then compute rank product
4. Identify genes with highest rank product

Linking to Genes: PC 1-3

microRNA:	RNA-seq:	Methylation:
hsa.miR.1260	ENSG00000257093.2	cg24794228
hsa.miR.222	ENSG00000118785.9	cg05585513
hsa.miR.95	ENSG00000151552.7	cg18329187
hsa.miR.132	ENSG00000164402.9	cg13580710
hsa.miR.200a	ENSG00000166535.15	cg11532431

Next Steps:

- Deep dive into identified genes: explore literature to see if there is evidence of relationships to AD
- Observe characteristics of the subset of subjects that appears to have strong connections to AD
- View patterns in genetic information between the identified clusters

References:

Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(12), 4164–4169. https://doi.org/10.1073/pnas.0308531101 (2004).

McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. https://arxiv.org/abs/1802.03426 (2018).

Tang, T. & Allen, G. Integrated Principal Components Analysis. https://arxiv.org/abs/1810.00832 (2018).